



PATENT ABSTRACTS OF JAPAN

(11) Publication number: **06261755 A** /(43) Date of publication of application: **20 . 09 . 94**

(51) Int. Cl

C12N 15/09(21) Application number: **05050303**(71) Applicant: **TOSHIBA CORP**(22) Date of filing: **11 . 03 . 93**(72) Inventor: **NAKAE HIROKI**(54) **GENE ANALYSIS**

(57) Abstract:

PURPOSE: To give a numerical value to a gene quality relating to a usage of codon in an arbitrary region of the gene by weighting a codon coding amino acids and calculating weighted average of all codon used for the region to be gene analyzed.

CONSTITUTION: In plural codons corresponding to each of all amino acids, when 1 is given to a codon used in high frequency and 0 to one used in low frequency and

this weighting is applied to an objective gene to be analyzed to calculate the weighted average, it is able to judge compatibility of the gene to be analyzed in case when a microorganism which is a standard of the weighting is used as a host. When the weighting is carried out, for plural codons that code amino acids, according to a number of GC pairs which are included in the codon, it is able to give a numerical value to a gene quality in a given region of the gene by calculating the weighted average used in the region.

COPYRIGHT: (C)1994,JPO&Japio

(19)日本国特許庁(J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平6-261755

(43)公開日 平成6年(1994)9月20日

(51)Int.Cl. ⁵	識別記号	庁内整理番号	F I	技術表示箇所
C 1 2 N 15/09		9050-4B	C 1 2 N 15/ 00	A

審査請求 未請求 請求項の数 1 O L (全 6 頁)

(21)出願番号 特願平5-50303

(22)出願日 平成5年(1993)3月11日

(71)出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72)発明者 中江 裕樹

神奈川県川崎市幸区小向東芝町1番地 株

式会社東芝研究開発センター内

(74)代理人 弁理士 鈴江 武彦

(54)【発明の名称】 遺伝子解析方法

(57)【要約】

【構成】アミノ酸をコードする各コドンに重み付けをし、遺伝子全体を含む、解析しようとする領域に使用される全てのコドンに付与された重みの平均を算出する。各コドンに付与される重みは任意に設定する。

【効果】遺伝子全体を含む、任意の領域のコドン使用に関する性質を数値化することができ、宿主とこの宿主に導入しようとする遺伝子との適合性、解析対象遺伝子の耐熱性を容易に判定することができる。

【特許請求の範囲】

【請求項1】 アミノ酸をコードするコドンに重み付けをし、遺伝子の解析しようとする領域に使用される全てのコドンに付与された重みの平均を計算することを特徴とする遺伝子解析方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 この発明は、遺伝子の性質を解析することが可能な遺伝子解析方法に関する。

【0002】

【従来の技術】 mRNAからのタンパク質の翻訳においては、複数のコドンが同一のアミノ酸に対応する場合があることが知られており、各コドンの使用頻度は種によって異なっている。

【0003】 近年急速に発展する遺伝子工学においては、特定の細胞、微生物等の宿主に他の遺伝子を導入して、宿主内で導入した遺伝子を発現させることが広く行なわれている。この際、宿主と導入した遺伝子とで各コドンの使用頻度が似ている場合には導入した遺伝子が発現し易くなる。逆に、使用頻度が大きく異なると導入した遺伝子の発現は困難になる。

【0004】 また、遺伝子DNAは、アデニン（A）とチミン（T）およびグアニン（G）とシトシン（C）とがペアをなして二本鎖を形成しているが、A-TペアよりもG-Cペアのほうが結合力が強く、遺伝子全体を見た場合G-Cペアの含量が高いほうが耐熱性に優れていることが知られている。

【0005】 したがって、遺伝子を解析してコドンの使用頻度やG-Cペアの含量を測定することにより、導入しようとする遺伝子の宿主適合性や耐熱性を予め知ることができる。また、既に存在する遺伝子のみではなく、遺伝子を設計する際にもこの手法は有用である。

【0006】 従来、遺伝子を解析するためのソフトウェアとしては、GENETYX（ソフトウェア開発社）や、DNASIS（日立ソフトウェアエンジニアリング社）などが知られている。これら遺伝子解析ソフトウェアには、遺伝子の使用コドンの解析機能として、各コドンの使用頻度や、遺伝子全体に亘って、コドンの1文字目、2文字目、3文字目、それぞれのG-C含量を計算する機能が備わっている。また、使用コドンの片寄りを表すパラメータとしては、従来より、RSCU値（Sharp, P.M. et al. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes, Nucleic Acids Research, 14, 5125-5143）が知られている（例えば、日本分子生物学会編、「DNAの構造と動態」、丸善、pp.219-220）。このパラメータは、コドンの使用回数を、同一のアミノ酸をコードするコドンの間で計算した平均使用回数で割った値で表される。

【0007】 さらに、「遺伝子全体のG-C含量」と「同

ジアミノ酸をコードするコドンのうち、どの程度多くG-Cペアを含むコドンを選択しているかの度合い」との間には、密接な関係があり、遺伝子の性質を調べるために、しばしばこの「同じアミノ酸をコードするコドンのうち、どの程度多くG-Cペアを含むコドンを選択しているかの度合い」を知る必要が生じていた。この目的のため、従来は、最もアミノ酸の選択に影響が少ないコドンの3文字目のG-C含量で、上述の度合いを表現している。

10 【0008】

【発明が解決しようとする課題】 しかし、前述の遺伝子解析ソフトウェアにより解析されるコドンの使用頻度並びにRSCU値は、個々のコドンに対するパラメータであり、遺伝子全体に対するパラメータではない。したがって、個々のコドンに関しては使用頻度の片寄りを知ることができるが、遺伝子全体に対する値にはなり得ない。

20 【0009】 また、コドンの3文字目のG-C含量を決定する方法では、ロイシンをコードするコドンであるUUGとCUUとのように、3文字目がG-Cペアだからといって、全体のG-Cペアの個数が必ずしも多くない場合もある。加えて、メチオニンやトリプトファンなどのように1つのコドンのみでコードされているアミノ酸の場合が考慮できないなどの欠点があった。

【0010】 このように、従来から知られている遺伝子解析方法は、コドンの使用に関しては個々のコドンについて解析するものであり、遺伝子全体の性質を反映するものではない。

30 【0011】 また、遺伝子の性質の解析において、しばしば「同じアミノ酸をコードするコドンのうち、どの程度多くG-Cペアを含むコドンを選択しているかの度合い」を知る必要が生じているにも関わらず、それを正確に表す値は従来から知られておらず、したがってそれを計算して遺伝子の性質を解析する遺伝子解析方法も存在しない。この発明は、遺伝子全体を含む任意の領域のコドン使用に関する性質を数値化することが可能な遺伝子解析方法を提供することを目的とする。

【0012】

40 【課題を解決するための手段】 本発明者は、上記事情に鑑み、鋭意研究の結果、この発明を完成するに至った。すなわち、この発明の遺伝子解析方法は、アミノ酸をコードするコドンに重み付けをし、遺伝子の解析しようとする領域に使用される全てのコドンに付与された重みの平均を計算することを特徴とする。この発明の遺伝子解析方法においては、各コドンに付与する重みは任意に設定することができる。

【0013】 例えば、特定の微生物でのmRNAからのタンパク質の翻訳過程において、全アミノ酸それぞれに対応する複数のコドンのうち、使用頻度の高いものに1、低いものに0を付与する。この重み付けを解析対象

遺伝子に適用して重みの平均を算出することにより、重み付けの基準となった微生物を宿主とした場合の解析対象遺伝子の適合性を判定することができる。すなわち、解析対象遺伝子の重みの平均が1である場合にはその遺伝子と微生物との両者で使用頻度が高いコドンが共通であることを意味する。逆に、重みの平均が0である場合には、解析対象遺伝子が、基準となった微生物では使用頻度が低いコドンのみを高い頻度で使用することを意味する。したがって、重みの平均が1に近いほどその微生物が解析対象遺伝子の宿主として適していることを示している。

【0014】また、あるアミノ酸をコードする複数のコドンに対して、そのコドンを含むGCペアの数に応じて重み付けをして、ある遺伝子の一定区域に対して使用されるコドンの重みの平均を計算し、当該遺伝子の当該区域の性質を数値化することもできる。この発明においては、このようにして数値化された性質をGC選択率と称する。

【0015】GC選択率を算出するための重み付けの例を挙げると、フェニルアラニンをコードするコドンUUUに-1、UUCに1、ロイシンをコードするコドンUUAに-1、UUGに0、CUUに0、CUCに1、CUAに0、CUGに1、イソロイシンをコードするコドンAUUに-1、AUCに1、AUAに-1、メチオニンをコードするコドンAUGに0、バリンをコードするコドンGUUに-1、GUCに1、GUAに-1、GUGに1、セリンをコードするコドンUCUに-1、UCCに1、UCAに-1、UCGに1、AGUに-1、AGCに1、プロリンをコードするコドンCCUに-1、CCCに1、CCAに-1、CCGに1、スレオニンをコードするコドンACUに-1、ACCに1、ACAに-1、ACGに1、アラニンをコードするコドンGCUに-1、GCCに1、GCAに-1、GCGに1、チロシンをコードするコドンUAUに-1、UACに1、ヒスチジンをコードするコドンCAUに-1、CACに1、CAAに-1、CAGに1、アスパラギンをコードするコドンAAUに-1、AACに1、リジンをコードするコドンAAAに-1、AAGに1、アスパ*

* ラギン酸をコードするコドンCAUに-1、GACに1、グルタミン酸をコードするコドンGAAに-1、GAGに1、システインをコードするコドンUGUに-1、UGCに1、トリプトファンをコードするコドンUGGに0、アルギニンをコードするコドンCGUに0、CGCに1、CGAに0、CGGに1、AGAに-1、AGGに0、グリシンをコードするコドンGGUに-1、GGCに1、GGAに-1、GGGに1という重みを付けることができる。

10 【0016】さらに、このGC選択率を利用して、特定遺伝子全体の性質をより詳細に調べることができる。すなわち、所定数のコドンを1単位とし、特定の遺伝子の末端から末端へ、この単位毎のGC選択率を遺伝子全体に亘って算出することにより、その遺伝子内のGC選択率の分布を計算することが可能となる。具体的な例を挙げると、10コドンを1単位とし、ある遺伝子の第1コドンから第10コドンまでのGC選択率を算出し、次いで第2コドンから第11コドン、第3コドンから第12コドンと同様の計算を遺伝子全体に亘って繰り返す。

20 【0017】

【作用】アミノ酸をコードする複数のコドンのそれぞれに重みを付け、遺伝子全体もしくは所望の領域についての重みの平均を算出することにより、当該遺伝子の性質を数値化し、重み付けの基準となった性質に対する当該遺伝子の位置付けを行なうことができる。

【0018】

【実施例】

実施例1

30 【0019】この実施例では、ホヤとマウスのエンタクチンを例にとり、大腸菌を宿主ベクター系とした場合どちらが発現させ易いかを知るために、どれだけ大腸菌にとって好ましいコドンを使用しているかを解析した。表1に各アミノ酸に対応するコドンと、この実施例のために設定した各コドンの重み、並びにホヤとマウスのエンタクチンにおける各コドンの使用頻度を示す。

【0020】

【表1】

5					6			
aa	c	w1	w2	w3	E. c.	ホヤ	マウス	ヒト
Phe	TTT	0	0	-1	2	21	15	13
	TTT	1	1	-1	17	31	35	40
Leu	TTT	0	0	-1	1	5	2	3
	TTT	1	1	-1	1	20	9	10
	CTT	0	0	0	1	12	10	11
	CTT	1	1	0	1	5	14	14
	CTA	0	0	0	0	8	8	3
	CTG	1	1	1	24	9	42	40
Ile	ATT	0	0	-1	0	31	23	22
	ATC	1	1	-1	10	21	3	30
	ATA	0	0	-1	0	11	3	4
Met	ATG	1	1	0	4	21	16	8
Val	GTT	1	0	-1	12	29	88	16
	GTC	0	1	-1	2	27	15	19
	GTA	1	0	-1	9	21	14	4
	GTG	1	1	-1	2	24	46	47
Ser	TCT	1	0	-1	6	12	15	13
	TCC	1	1	-1	8	16	25	31
	TCA	1	0	-1	0	16	1	10
	TCG	1	1	-1	0	10	3	3
	AGT	1	0	-1	1	12	5	12
Pro	AGC	1	1	-1	2	11	34	24
	CCT	0	0	-1	0	19	32	24
	CCC	0	0	-1	0	9	35	36
	CCA	0	0	-1	3	28	18	15
Thr	CCG	1	1	-1	1	12	4	12
	ACT	1	0	-1	12	26	16	17
	ACC	1	1	-1	12	16	39	40
	ACA	0	0	-1	0	27	22	16
Ala	ACG	0	1	-1	0	7	9	23
	GCT	1	0	-1	18	22	18	22
	GCC	0	1	-1	0	9	32	23
	GCA	1	0	-1	8	30	18	17
	GCG	1	1	-1	3	6	6	17
Tyr	TAT	0	0	-1	5	21	25	15
	TAC	1	1	-1	24	21	22	29
His	CAT	1	0	-1	0	11	10	9
	CAC	1	1	-1	1	8	30	25
Gln	CAA	0	0	-1	1	27	7	11
	CAG	1	1	-1	20	15	50	47
Asn	AAT	0	0	-1	0	33	21	18
	AAC	1	1	-1	32	21	25	27
Lys	AAA	1	0	-1	17	21	7	7
	AAG	0	1	-1	0	16	22	22
Asp	GAT	1	0	-1	9	44	36	25
	GAC	1	1	-1	23	38	38	41
Glu	GAA	1	0	-1	11	66	32	27
	GAG	0	1	-1	0	20	39	46
Cys	TGT	1	0	-1	0	16	22	17
	TGC	1	1	-1	0	20	28	33
Trp	TGG	1	1	0	4	11	15	13
Arg	CGT	1	0	0	12	13	5	3
	CGC	1	1	0	1	9	11	19
	CGA	0	0	0	0	14	13	10
	CGG	0	0	1	0	5	9	14
	AGA	0	0	-1	0	24	16	12
	AGG	0	1	0	0	4	14	15
Gly	GGT	1	0	-1	29	31	17	11
	GGC	1	1	-1	19	17	42	42
	GGA	0	0	-1	0	39	30	25
	GGG	0	1	1	0	12	28	35
					367	1161	1245	1247

【0021】表1において、aa はアミノ酸の種類を示し、Phe はフェニルアラニン、Leuはロイシン、Ile はイソロイシン、Met はメチオニン、Val はバリン、Ser はセリン、Pro はプロリン、Thr はスレオニン、Ala はアラニン、Tyr はチロシン、His はヒスチジン、Gln はグルタミン、Asn はアスパラギン、Lys はリジン、Asp はアスパラギン酸、Glu はグルタミン酸、Cys はシステイン、Trp はトリプトファン、Arg はアルギニン、Gly はグリシンをそれぞれ表わす。また、Cはそれぞれのアミノ酸をコードするコドンを示し、ここでは、mRNAではなくDNA上の配列で示してある。Cの列のA、T、G、Cは、それぞれDNAを構成する塩基を示しており、Aはアデニン、Tはチミン、Gはグアニン、Cは

シトシンをそれぞれ表わしている。

【0022】この表でw1は、大腸菌での使用頻度の低いコドンには0を、それ以外のコドンには1を設定した重みである。また、E. c. は大腸菌ompC遺伝子の、ホヤ、マウス、ヒトは、それぞれホヤ、マウス、ヒトのエンタクチン遺伝子のアミノ酸コード領域のコドン使用頻度を示している。

【0023】この表から、コドンの使用回数とw1の重みをかけ、その合計を全コドン数で割る。すなわち、それぞれの遺伝子において重みの平均を計算すると、E. c. では0.95、ホヤでは0.60、マウスでは0.62であった。この平均は、大腸菌の遺伝子に近ければ、1に近づいたため、大腸菌の遺伝子であるE. c. は、0.95という非常に1

に近い値を示している。一方、ホヤとマウスのエンタクチンの遺伝子は、ホヤが0.60、マウスが0.62であり、大きな違いはないが、若干マウスのほうが大きい値を示した。この値はどれだけ大腸菌にとって好ましいコドンを使用しているかを示しているため、この値を用いれば、どれだけ大腸菌で発現させ易い遺伝子かということを知ることができる。したがって、大腸菌を宿主ベクター系として用いる場合には、ホヤエンタクチン遺伝子に比べてマウスエンタクチン遺伝子のほうが若干発現させ易い遺伝子であることが分かる。

【0024】また、表1のw2は、コドンの3番目の塩基がAもしくはTの場合には0を、GもしくはCである場合には1を設定した重みである。このように、コドンの3番目の塩基に応じて重みを設定すれば、その重みの平均値から3dコドンのGC含量を求めることも可能である。計算の結果、3dコドンのGC含量はホヤで40%、マウスで61%であった。

実施例2

【0025】この実施例では、ホヤおよびマウスのエンタクチン遺伝子が、「同じアミノ酸をコードするコドンのうち、どの程度多くのGCペアを含むコドンを選択しているか」を明らかにした。そのためまず、アミノ酸をコードするコドンに、次式に従って重みをつけた。

【0026】

【数1】

$$w_{ij} = \frac{N_{ij} - N_{\min i}}{N_{\max i} - N_{\min i}} - 1$$

【0027】ここで、 w_{ij} はアミノ酸iをコードするコドンjに対する重み、 N_{ij} はアミノ酸iをコードするコドンjに含まれるGCペアの数、 $N_{\min i}$ はアミノ酸iをコードするコドンのうちGCペアの最も少ないコドンに含まれるGCペアの数、および $N_{\max i}$ はアミノ酸iをコードするコドンのうち最も多くGCペアを含むコドンに含まれるGCペアの数をそれぞれ示す。ただしアミノ酸iには、1種類のコドンでコードされているメチオニンとトリプトファンは含めず、これら2つのアミノ酸の重みは常に0とする。このようにして付けられた重みを基にしてGC選択率(GCSR)を次式によって計算する。

【0028】

【数2】

$$GCSR = \frac{1}{n} \cdot \sum (w_{ij} \cdot X_{ij})$$

ここで、nはコドンの合計数、 X_{ij} はアミノ酸iをコードするコドンjの使用回数を示している。

【0029】表1に示す各コドンの使用回数からホヤとマウスのGCSRを計算すると、それぞれ -0.19、0.21

となる。したがって、マウスは、ホヤに比べると、同じアミノ酸をコードするコドンのうち多くのGCペアを含むコドンをより多く選択しているということが分かる。

実施例3

【0030】この実施例では、ホヤ並びにヒトのエンタクチン遺伝子が、「同じアミノ酸をコードするコドンのうち、どの程度多くのGCペアを含むコドンを選択しているか」を明らかにした。そのためまず、アミノ酸をコードする各コドンに表1のw3に示すように重みをつけ、次いで実施例2に定義するGC選択率を数2に示す式に従って算出した。すなわち、コドンの使用回数とw3の重みをかけ、その合計を全コドン数で割り、それぞれの遺伝子において重みの平均を計算した。その結果、ホヤでは実施例2と同様に -0.19、ヒトでは0.30であった。したがって、ホヤに比べヒトでは、同じアミノ酸をコードするコドンのうち多くのGCペアを含むコドンをより多く選択しているということが分かる。

実施例4

この実施例では、この発明の遺伝子解析方法を用いて、ホヤエンタクチン遺伝子内のGC選択率の分布を明らかにした。

【0031】各アミノ酸をコードするコドンの重みは、実施例3と同様、表1のw3に示す値を用い、10コドンを1単位としてGC選択率を遺伝子全体に亘って求めた。すなわち、第1コドンから第10コドンまで、第2コドンから第11コドンまで、第3コドンから第12コドンまで、第4コドンから第13コドンまでというように、ホヤエンタクチン遺伝子の末端から10コドンずつを1単位としてGC選択率を求めていき、最後に第1152コドンから他方の末端である第1161コドンまでのGC選択率を求めた。その結果を図1に示す。この図からは、全体に亘ってGC選択率が低く、また2か所に特にGC選択率の低い部分があるというホヤエンタクチン遺伝子の性質が明らかである。

【0032】このように、ホヤエンタクチン遺伝子に対して、予め定めた10コドンの長さ範囲のGC選択率をホヤエンタクチン遺伝子全体に亘って計算し、その遺伝子内のGC選択率の分布を求めることによって、遺伝子全体の性質を調べることができる。

【0033】

【発明の効果】以上のように、この発明の遺伝子解析方法によると、遺伝子全体を含む任意の領域のコドン使用に関する性質を数値化することができる。

【0034】この発明の遺伝子解析方法によって数値化される性質は、各コドンにどのような重みを付与するかにより任意に設定することができる。例えば、特定の微生物のタンパク合成系において使用頻度の高いコドンにより重い重みを付与することにより、その微生物を宿主とした場合の解析対象遺伝子の適合性を判定することができる。また、GCペアを含有するコドンにより重い重

みを付与することにより、解析対象遺伝子がGCペアを含むコドンを用いる傾向を見ることができ、その結果、その遺伝子の耐熱性を判定することができる。さらに、この発明の遺伝子解析方法を利用することにより、遺伝子の設計を行なう際に、予め遺伝子の性質を予測するこ*

*とが可能となる。

【図面の簡単な説明】

【図1】この発明の遺伝子解析方法を用いて得られた、ホヤエンタクチン遺伝子におけるGC選択率の遺伝子上での分布を示すグラフ。

【図1】

